Correlation Between Suicide Rates and Socio-Economic Statistics

Evgeniy Ko, Chung Hyun Lee, Tek Acharya

Georgia State University
Data Mining
09 December 2020

Correlation Between Suicide Rates and Socio-Economic Statistics

Evgeniy Ko
Georgia State University

Chung Hyun Lee
Georgia State University

Tek Acharya
Georgia State University

Abstract

Suicide is a tragic event and is the cause of around 800,000 global deaths annually. In the United States annual suicide rates increased by 24% from 1999 to 2014, and in that time frame, the United State's GDP per capita increased from $35,514 to $55,048. Based off of a book published by the World Health Organization[2], which took data from 2012, regions with a lower-middle income population accounted for 35.4% of the global population, but accounted for 41.4% of global suicides, and has a suicide rate of 14.1%, which is the highest suicide rate across all levels of income. Low-income regions account for 12% of the global population and 10.2% of global suicides, and has a suicide rate of 13.4%, which is second to lower-middle-income regions. High-income regions account for 18.3% of the global population and 24.5% of global suicides, and has a suicide rate of 12.7%. Upper-middle-income regions account for 34.3% of the global population and 23.8% of global suicides, and has a suicide rate of 7.5% which is the lowest amongst all levels of income.

1 Introduction

With this knowledge, our group's motivation for this project is to see if we can draw a correlation between suicide rates and a countries' annual GDP. If we were to strengthen our understanding of why suicide rates increase, we might improve our understanding of how to decrease them as well. For this project we have decided to use a data set of socio-economic statistics of 101 countries from 1985 to 2015. We will conduct this project by applying three machine learning algorithms that will help us determine if a correlation exists. The first machine learning algorithm we want to apply is PCA. PCA is a machine learning algorithm which will find a principal component which is a combination of the initial variables which tries to explain as much of the variance in the data set as possible. The second algorithm we want to utilize for our project is k-means clustering. With k-means we hope to find common characteristics in clusters that might result in higher or lower suicide rates. The last algorithm we wish to apply is multiple linear regression. With this, we hope to find a regression line based off of each of our variables.

2 Data Analysis

	suicides_no	population	suicides/100k pop	gdp_for_year ($)	gdp_per_capita ($)
count	27820.000000	2.782000e+04	27820.000000	2.782000e+04	27820.000000
mean	242.574407	1.844794e+06	12.816097	4.455810e+11	16866.464414
std	902.047917	3.911779e+06	18.961511	1.453610e+12	18887.576472
min	0.000000	2.780000e+02	0.000000	4.691962e+07	251.000000
25%	3.000000	9.749850e+04	0.920000	8.985353e+09	3447.000000
50%	25.000000	4.301500e+05	5.990000	4.811469e+10	9372.000000
75%	131.000000	1.486143e+06	16.620000	2.602024e+11	24874.000000
max	22338.000000	4.380521e+07	224.970000	1.812071e+13	126352.000000

Figure 1: Description of the numerical variables

Our dataset consists of 27,820 data samples which displays information of each gender-age group for 101 different countries. The data set is composed of twelve columns, of these twelve columns, six of them are categorical: 'country', 'year', 'sex', 'age', 'country-year', and 'generation'; and six of them are numerical: 'suicides_no', 'population', 'suicides/100k pop', 'HDI for year', 'gdp_for_year ($)', and 'gdp_per_capita ($)'. 'HDI for year' is a variable that consists of mostly missing values, so we will omit this column instead of filling in the missing values. We remove the generation column because it is the same as our age group column.

	country	year	sex	age
count	27820.000000	27820.000000	27820.000000	27820.000000
mean	49.275270	16.258375	0.500000	2.499425
std	29.372538	8.469055	0.500009	1.708754
min	0.000000	0.000000	0.000000	0.000000
25%	24.000000	10.000000	0.000000	1.000000
50%	47.000000	17.000000	0.500000	2.000000
75%	74.000000	23.000000	1.000000	4.000000
max	100.000000	31.000000	1.000000	5.000000

Figure 2: Description of the categorical variables

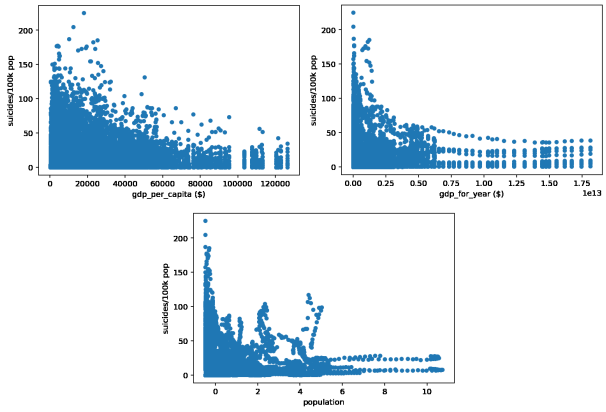

Figure 3: Suicides per capita plotted against GDP per capita, GDP per year, and population

We can also remove the country-year column, because we already have a column for country and year. The other columns do not need any adjustments since they do not have any missing values.

2.1 Preprocessing

For the preprocessing of our data, we adjusted most of our variables so that we can achieve more precise and accurate results from our algorithms. We preprocessed the 'country', 'year', 'sex', and 'age' variables by using the pandas cat.codes function, which gives numerical values to categorical variables. 'HDI for year' is a variable that consists of mostly missing values, so we omit this column instead of filling in the missing values with inaccurate data. We removed the generation column because when giving dummy values, it will be the exact same as our age group column. We can also remove the country-year column, because a country and year column already exist. Other than those columns, the other columns do not need any adjustments since they don not have any missing values.

2.2 Data Exploration

2.2.1 Visualization

Upon first glance of each of variable, we can identify several patterns in our data set. Based on the graphs in Figure 3, we can see that there could possibly be an exponential decay in suicides per capita, as a country's GDP per capita (top left), GDP per year (top right), and population increase (bottom). In Figure 4, we can notice that the age group with the highest rate of suicide are people above the age of 75, while the age group with the lowest rates of suicide are people between the ages of 5 and 14 (top left graph). We can also notice that males have a higher average suicide rate than females (top right graph). The last two graphs in Figure 4 shows the rate of suicides (bottom left) and the average GDP per capita (bottom right) from 1985 to 2016.

Understanding In 2005 the average GDP per capita of all 101 countries significantly rose, however, the rate of suicide in 2005 was low compared to the years before, and kept decreasing until 2015. Based off of this observation from our data set, it seems like GDP per capita does not have a strong correlation with suicides per capita. If there was a correlation, we would see an increase in suicide rates over time since the average GDP per capita of our 101 countries is significantly greater in 2016 than it was in 1985. However, these statistics do not account for the increase in cost of living and product pricing.

3 Related Work

An article[1] published in 2014 aimed to analyze how unemployment, GDP per capita, annual economic growth rate, and inflation affected number of deaths by suicide in men and women. The article found a strong correlation between annual economic growth and suicide rates in men, and a correlation with unemployment and suicide rates in women. they also identified an increase in suicide rates several months prior to the occurrence of an economic crisis. With this information, the researchers confirmed a relationship between the economic environment and suicide rates in Europe. However, no two countries in the world are identically

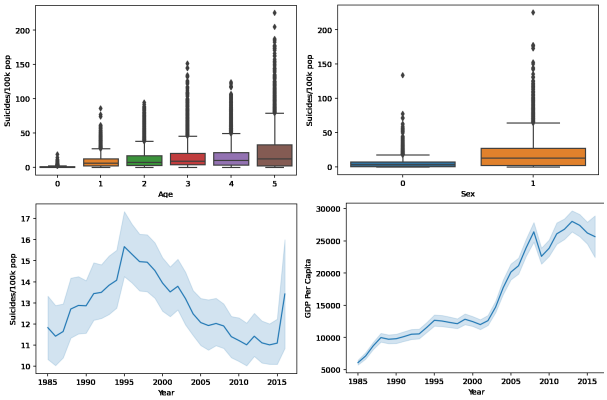

Figure 4: Suicides per capita plotted against age, sex, and year

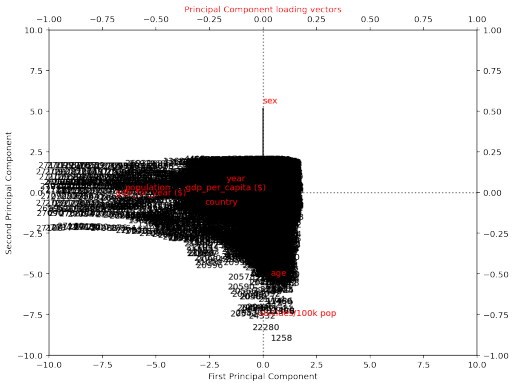

Figure 5: Our resulting PCA graph

the same, therefore, we cannot apply these results to countries outside of the European countries used in the research.

4 Algorithm Design

4.1 PCA

Principal Component Analysis (PCA) is a machine learning algorithm that is used to reduce the dimensionality of datasets to increase its interpretability. The PCA algorithm will produce a principal component which is the combination of the initial variables that best accounts for the shape of the data set. For this algorithm, we performed some or adjustments to our data set by removing the 'country-year' and 'generation' variables because they are duplicates to variables we already have. We scaled all of our data to improve the accuracy of our algorithm, the scale of our 'population', and GDP related variables are much greater than the rest of our variables, so we do not want small changes in those variables to affect the results more than if a change in another variable is greater in terms of its respective standard deviation.

Our results show that our first principal component explains 23% of the variance, and our second principal component explains 19% of the variance. The highest explained variance by two variables from our PCA algorithm is 42%. With such a low explained variance in our first principal component, we cannot say that there is a strong correlation between our variables. Using our first principal component would not give us accurate results when trying to find how our variables affect suicides per capita.

4.2 K-Means

K-Means clustering is a machine learning algorithm that identifies a k number of centroids, and then allocates every data point to the nearest cluster to help find underlying patterns. We used the elbow method to help find our k value. The elbow method uses k-means clustering on the data set with a range of values for k, we used a range of 1 to 10, and then computes the average score for each of these k values. Then you map the results on a graph that should resemble an arm, and the "elbow" of the graph is the best possible value of k. The k value that we found was 5.

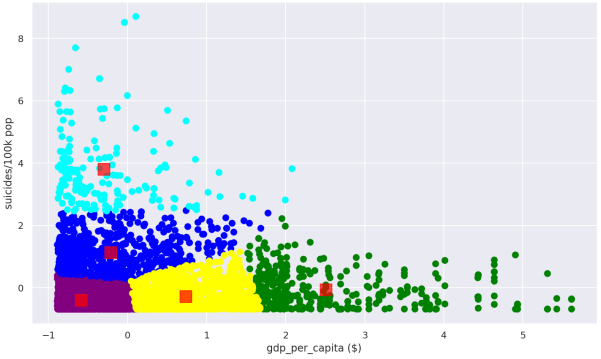

Figure 6: Our resulting k-means graph

After applying the algorithm to our data set with a k

value of 5, we were given the graph shown in Figure 6. From this graph, we identified that the purple, yellow, and green clusters were grouped together because they had low rates of suicide for their GDP per capita, while blue and cyan represented a higher rate of suicide for what their GDP per capita was. The k-means score that we got for our model was -918.91. The k-means score is a large number because there is a large variety in our data samples. The number of data samples that we have is much greater than the number of clusters we have, which is only 5 compared to the 27,820 samples in our data set.

4.3 Linear Regression

Linear regression is a machine learning algorithm that models a relationship between a dependent variable and one or more explanatory variables. The score from our PCA was too low, so we cannot use the principal components for our linear regression model. Instead, we are using multiple linear regression to evaluate all of our variables independently from each other.

4.3.1 GDP per Capita vs Suicides per Capita

In Figure 7, you can see our resulting linear regression model after applying the algorithm on our GDP per capita vs suicides per capita (left) and GDP per year vs suicides per capita graphs (right). For these graphs, we normalized the variables so the scale of our coefficent and intercept were not extraordinarily small. The model we received from applying linear regression on our GDP per year vs suicides per capita graph returned with a coefficient of 0.0136 and an intercept of -0.0045. The model received an R^2 score of $9.77 * 10^{-5}$. This means that there is no strong correlation between GDP per capita and suicides per capita. We tried to apply polynomial regression to see if we can achieve a better score, however, with the degree set to 3, we were only able to achieve a score of 0.0006, which is insufficient to say that there is a correlation between GDP per capita and suicides per capita. We expected to find better results from this model because of the trend observed from the graph. However, the reason we found this algorithm was unsuccessful is because the amount of data points on the left side of our graph far outweighs the amount of data points on the right side of our graph.

4.3.2 GDP per Year vs Suicides per Capita

The model we received from applying linear regression on our GDP per year vs suicides per capita graph (right) returned with a coefficient of 0.0187, and an intercept of 0.001. However, the R^2 score our model received when applying it to our data set was 0.0004,

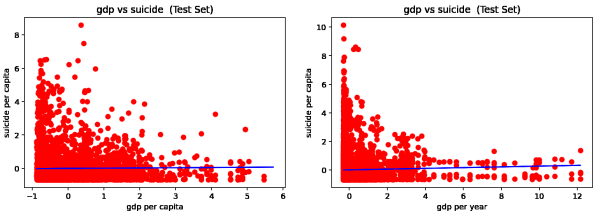

Figure 7: Linear regression performed on GDP per capita and GDP per year against suicides per capita

meaning that GDP per year and suicides per capita did not have a strong correlation. With such a low score, we decided to test if polynomial regression might produce a better score, since there does seem to be an exponential decay in the suicides per capita as GDP per year increases. However, even with a degree of 3, we were only able to get a score of 0.0016, only slightly better compared to our linear regression score, and still not sufficient enough to say that there is a correlation between GDP per year and suicides per capita.

4.3.3 Age vs Suicides per Capita

Based on the analyses we performed on the age graph in Figure 8, we predict that there is a correlation between age and suicides per capita. There are a lot of outliers in this graph so we decided to remove them by removing all data points that are 1 standard deviation away from the mean suicides per capita for their respective age group. The model we received from applying linear regression on our age vs suicides per capita training set (top left) in Figure 8 returned with a coefficient of 2.275 and an intercept of 2.1914. After applying it to our test set shown in Figure 8 (bottom left) the R^2 we received for this model was 0.1836, which is an improvement compared to when we did it without removing outliers by 0.05, and the highest score from all of our regression models, however, this score is also not good enough to say there is a strong correlation between age and suicides per capita. We also performed polynomial regression on this graph with a degree of 2 and 3, however, the best score we could get was 0.1989 with a degree of 3.

4.3.4 Year vs Suicides per Capita

The last variable we chose to apply our linear regression model to was our year vs suicides per capita graph. After applying the algorithm to this graph, it returned with the model shown in Figure 8 (right), and has a coefficient of -0.1167, an intercept of 14.8912, and an

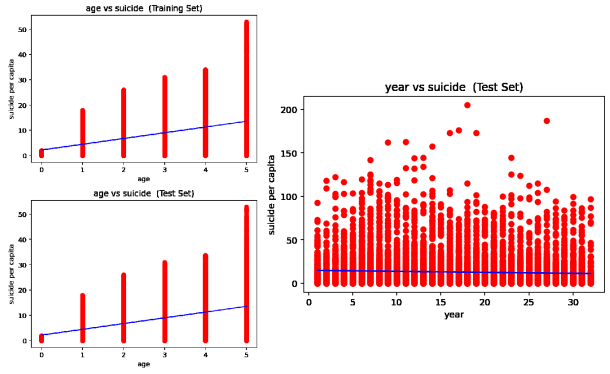

Figure 8: Linear regression performed on year and age against suicides per capita

R^2 score of 0.0026. After applying polynomial regression, we were able to achieve a score of 0.0041 which is still an inadequate R^2. These results are not sufficient enough to be able to conclude that there is a correlation between time and suicides per capita.

5 Conclusion

Based on the results of our machine learning algorithms, we cannot say that there is a correlation between the socio-economic statistics of a country, and suicides per capita. All of our machine learning algorithms returned with scores that tell us the models we created explain very little of the variance in our data set.

5.1 PCA

The PCA algorithm returned a first principal component that explained 23% of the variance in our data set. This explained variance score is too low for us to be able to reliably use the principal component and be able to determine a country's suicide rate based on their socio-economic statistics. Based on these results we would say that PCA was ineffective in reducing the amount of variables we can work with.

5.2 K-Means

After applying the k-means clustering algorithm on our data set with a k value of 5, we were given a score of -919.91 which is not a good score. Because there are thousands of data points in a single cluster, we cannot identify any patters in data points from the same centroid. K-means was unsuccessful because our data set is significantly larger than the number of clusters we have, therefore the sum of all distances between data samples is too large.

5.3 Linear Regression

The GDP per capita, GDP per year, and year vs suicides per capita models we got from applying linear regression returned with an R^2 below .01. R^2 is a relevant statistic to use when evaluating the performance of our linear regression models because it represents the proportion of the variance of a dependent variable that is explained by an independent variable. However, we were able to receive an R^2 of 0.1836 from our age vs suicides per capita model. This means that the age of a person explains 18.36% of the variance shown in suicides per capita in that respective age group. We also tried to apply polynomial regression to our data set to see if we can obtain a more accurate model. The most significant improvement was for age, increasing by 1.5% to 19.9%. However, this model also does not explain enough of the variance for us to say there is a strong correlation between age and suicides per capita.

References

References

[1] Pavlos N. Theodorakis Ad J. F. M. Kerkhof Alvydas Navickas Cyril Ho schl Dusica Lecic-Tosevski Eliot Sorel Elmars Rancans Eva Palova Georg Juckel Goran Isacsson Helena Korosec Jagodic Ileana Botezat-Antonescu Ingeborg Warnke Janusz Rybakowski Jean Michel Azorin John Cookson John Waddington Peter Pregelj Koen Demyttenaere Luchezar G. Hranov Lidija Injac Stevovic Lucas Pezawas Marc Adida Maria Luisa Figuera Maurizio Pompili Miro Jakovljevic Monica Vichi Giulio Perugi Ole Andreassen Olivera Vukovic Paraskevi Mavrogiorgou Peeter Varnik Per Bech Peter Dome Petr Winkler Raimo K. R. Salokangas Tiina From Vita Danileviciute Xenia Gonda Zoltan Rihmer Jonas Forsman Benhalima Anne Grady Anne Katrine Kloster Leadholm Susan Soendergaard Carlos Nordt Konstantinos N. Fountoulakis, Wolfram Kawohl and Juan Lopez-Ibor. The british journal of psychiatry.

[2] World Health Organization. Global epidemiology of suicide and suicide attempts. *Preventing Suicide - A Global Imperative*, 1:24, 2014.